How Can Linguistics Help The Structuring of A Multidisciplinary Neo-Domain Such As Exobiology?

Anne Condamines

How Can Linguistics Help The Structuring Of A Multidisciplinary Neo-Domain Such As Exobiology ?[iii]

A. Condamines

CLLE-ERSS, UMR 5263, CNRS and University of Toulouse

Abstract. This chapter examines the possibility for linguistics to propose results from a corpus of texts analyse in order to contribute to stabilise definitions within exobiology. By using clues provided by tools, linguists build interpretation of contexts and try to show how the meaning is built taking into account the different points of view of the disciplines implied in exobiology. Examples of such interpretation are proposed.

1 Introduction

This chapter tries to show how corpus linguistics may help the structuring of a new and interdisciplinary domain such as exobiology. Since a few years, the linguistics has developed methods in order to study corpus of texts as systematically as possible. It is now easy to identify the most specific words in a corpus and, by using its contexts, to better understand their meaning. But exploring a corpus leads sometimes to discover that variations of meaning may appear for a same word in different parts of the corpus. Such phenomena infirm the hypothesis of a stable sense even within a scientific discipline. So, many linguists think that meaning is not given once and for all but that it has to be constructed and, if the aim is to define a concept, it is necessary to take into account all the variations of meaning appearing in a corpus. This point of view joins the one of most of epistemologists concerning the creation of concepts. In such a discipline as Exobiology, originate from several disciplines having they own points of view, we think that linguistics may help to identify these points of view and contribute to the definition of concepts. The second part presents the characteristics of exobiology from a linguistic point of view and lists the relevant elements of corpus linguistics for studying a multidisciplinary corpus. The third part shows how it is possible to build and analyze a corpus of exobiology in order to light its categories of semantic phenomena. The chapter is closed with the study of examples of these categories.

.

2 Characteristics of exobiology and linguistics for specialized corpora

This part presents the epistemological context of the study. It focuses on characteristics of exobiology on one hand and of corpus linguistics for specialized texts on the other. Taking into account these characteristics, it presents the way corpus linguistics can analyse multidisciplinary corpus and spot differences and similarities between the four disciplines implied in exobiology.

2.1 Exobiology

From the point of view of what interests the linguistics, exobiology presents two main characteristics. First, it is an emerging discipline and then, it is a multidisciplinary discipline. These two points are very important and they entail several consequences:

- This situation may generate complementary or conflicting viewpoints, more precisely, regarding lexicon, disciplines may either use words with a similar meaning or not.

- The concepts are in evolution. When experts from different fields agree to confront their point of view, it is clear that concepts are going to change under the influence of the multidisciplinarity.

- Definitions are not stable. This is a consequence of the evolution of the concepts: meanings themselves are fluctuating and, at least temporarily, definitions are destabilised.

This situation seems very close to the one described by Kuhn as a change of paradigm [1]. Nevertheless, there is perhaps a difference which is that with exobiology, most of the experts from each domain implied in the construction of exobiology are aware that it is necessary to enlighten collectively a similar object: extraterrestrial life.
From a linguistic point of view, the study of exobiology is really interesting because it is possible to analyse a discipline at the time that it is made up. The issue of building definitions in such a situation enables linguists to ask the role of definition both as a way to constitute referents for a discipline and also as a risk to limit the creation. As A. Rey notes [2], in the word *definition* as well as the word *term,* there is the meaning of limitation (finite). When you confine a term, you stabilize its meaning but you also limit variations in use. So building definitions is possible only when a high level of consensus is reached among the different experts. As says G. Fourez [3]:
« Les pratiques interdisciplinaires peuvent être considérées comme des négociations entre des points de vue et des intérêts différents, dans un contexte et selon un projet »[a].

2.2 Corpus Linguistics

First of all, here is a definition of a corpus : " a subset of [Electronic Text Library] built according to explicit design criteria for a specific purpose […] " [4]. We will see that, in our case, the specific purpose will be the spotting of differences between words and terms use in texts from different disciplines.
Here are the four characteristics of corpus linguistics ([5], [6])

a) Corpus linguistics needs necessarily a tool-helped approach.

The corpora analysed by linguists (even in specialized field) may contain several hundreds of thousands of word forms. So it may be very difficult to study and memorize all the contexts of a word. Moreover, Natural language processing tools propose very sophisticated clues giving an original enlightenment on linguistic phenomena (see below).

b) Corpus linguistics is doubly situated.

This point concerns the fact that in order to understand how the meaning may be apprehended, it is necessary to take into account both the situation in which a text has been written (by who (in our case, the discipline of the expert), the date of the writting ….) and the aim for which the meaning has to be built (in our case, for contributing to built a common point of view).

c) Corpus linguistics is often based on comparison of phenomena between several parts of the corpus.

A good way for understanding the meaning of the words is to compare them and their functioning within several sub-corpora. The organization of a corpus between sub-corpora plays a very important rule in such an approach because differences between meanings will be decided regarding the point of view that has been chosen for the organization

d) Corpus linguistics needs interpretation.

a

Starting from the premise that a meaning is not pre-existing but is built and that tools give clues, the linguists'role is to propose an interpretation of linguistic phenomena, taking these clues into account (and, in some cases, invalidate these clues).

Generally speaking, tools can provide three kinds of clues.
The first one concerns the quantitative aspect of the lexicon. At the moment, this point is the more developed in Natural Language Processing (NLP) which implements statistical methods in order to spot the words or patterns the more specific in a corpus. This clue is very often combined with the two others.
The second clue enables to compare forms variation. For example, with such a clue, it is possible to show that a name is systematically used with an adjective in a sub-corpus whereas it is not the case in another one.
The third clue is also very often used in NLP. It concerns a distributional point of view, developed by Harris from a mathematical point of view [7] and by Firth from a more sociological point of view [8]. As the latter said: *You shall know a word by the* company *it keeps*. From a semantic viewpoint, this has different consequences among which the fact that words meaning can be acquired only by using the contexts in which they appear (or their "distribution") and the fact that two words will be synonyms if they appear in the same contexts. As we will see in part XX, this clue is very useful in specialized domains [9], [10], [11]. But the main issue in such a distributional approach is the problem of contexts similarity. It is rare to find two different words in exactly similar contexts. Many times, it is necessary to decide if two contexts are similar or not, that is to say if they may be categorize as belonging to the same semantic category or not. This process of categorization concerns the first stage of the interpretation.

Finally, there is a main specificity in corpus linguistics for specialized corpus : linguists are not experts in the domain concerned by the corpus. So in such kinds of study, we can speak of "co-construction" of an interpretation as the result of two types of expertise: the one of the domain experts and the one of the language experts (linguists).

As many sciences philosophers [12], when adopting corpus linguistics methodology, most of linguists think that meaning is not stable and given once for quite but rather that it is built ([13], [14]), .
« The language of science demonstrates rather convincingly how language does not simply correspond to, reflect or describe human experience; rather, it interprets, or, as we prefer to say "construes" it. A scientific theory is a linguistic construal of experience" [9].
"Whatever reality may mean, it always correspond to an active intellectual construction. The description presented by science can no longer be disentangled from [scientists'] questioning activity" ([15])

This view is opposed to a traditional point of view on terminology focusing mainly on standardization :
"The new socio-cognitive theory of Terminology emphasises that Terminology should not be uniquely oriented towards standardisation and it questions the validity of objectivism as the theoretical underpinning of terminology". ([16])

We will see in the third part how corpus linguistics applied on specialized corpora may propose some elements in order to help this building.

Finally, we can resume this part saying that the analyze of an emerging discipline such as exobiology studying texts belonging to this neo-discipline needs to adopt a constuctivist point of view, sharing both by some philosophers of sciences and corpus linguists.

3 Corpus linguistics on the corpus OF EXOBIOLOGY

This part describes how corpus linguistics methods may be applied upon exobiology texts. By interpreting clues from tools and comparing texts from the four mains disciplines concerned by exobiology, it is possible to describe lexical and semantic phenomena and use them to spot both similarities and differences among the four disciplines [17].

3.1 The Corpus of exobiology

The corpus of exobiology has been built by Nathalie Dehaut (PhD student) with the help of M. Gargaud (lab. d'Astrophyisque de Bordeaux). The main difficulty was that there is no text really belonging to exobiology since the discipline is under construction. But there are some situations of communication which may be considered as representative of this construction. This is the case with schools of exobiology organised by CNRS (Centre National de la Recherche Scientifique) where experts of the main disciplines constituting exobiology intervene in order to present to advanced students or even specialised colleagues from other disciplines the main issues of their originate domain, linked with exobiology. This is a well known communicative situation (close to vulgarisation) well adapted to linguistic exploration. The corpus contains the two books edited after these summer schools : *L'environnement de la Terre Primitive et l'Origine de la Vie (2001)* et *Les traces du vivant et l'origine de la Vie (2003)*. These two books are written in French. This could be considered as a problem because the creation within exobiology concerns the international scientific community who writes mostly in english. But, on the one hand, it is easier for us who are French linguists to precisely analyse texts in our language and on the other, it is well known that, concerning the concepts organisation, there is no crucial difference within the same domain between the languages.

The corpus has been organised between four corpora, taking into account the scientific origin of the 38 authors. Here is the final organisation:

- Astronomy : 12 papers, 88,815 words
- Biology : 8 papers, 65, 589 words
- Chimistry : 8 papers, 80, 190 words
- Géology : 8 papers, 77, 010 words

3.2 Types of observed linguistic phenomena

Clues provided by tools have been interpreted in order to characterize lexical, semantic and discursive phenomena identified within the corpus.

3.2.1 Quantitative results

From a quantitative point of view, here are the most frequent words and their repartition in each domain. The table 1. shows the most present forms in the all corpus with their frequency (per thousand of words) in each sub-corpus.

Table 1. Most present forms in the all corpus

	Astro	Bio	Chem	Geo
Atmosphère	5,39	0,46	1,80	6,18
Eau	3,83	4,05	1,26	4,13
Temperature	2,57	2,67	1,13	4,43
Planète	5,26	0,61	0,37	3,24
Acide	0,45	1,83	6,42	0,18
Vie	2	2,65	1,73	2,19
Formation	0,72	0,77	2,43	2,74
Molécule	0,2	2,11	3	0,97

Three main comments may be made from this table. First, *acide* may be either a noun, either an adjective. Secondly, except *formation* which seems very general, the seven others words seem to be very linked to exobiology. Finally, half of the eight most frequent words of the corpus are provided by Geobiology. Astrobiology and Chemistry provide only one term each. But these comments should be taken with precaution because they concern only eight words. And, especially, what is important to note is that, within specialized corpora, if terms are nouns rather than verbs, they are mainly compound nouns.
Concerning *acide*, as noun, the adjective most present in each sub-corpus is *aminé*. It is in the chemistry corpus that possible adjectives are the most varied: more than 30 kinds. But *acide* may also be an adjective. For three of the corpus (Geobiology, chemistry and biology) it is the nouns *ph* that precedes *acide* most time. The other most frequent nouns are: for asronomy : *function* and *nuage*, for Geology: *roche* and *océan*, for chemistry: *milieu* and *environement* and for biology: *fonction* and *rivière*.
Two corpora present specificity in their use of *acide*. In Astronomy : *adsorption des acides* and in Chemistry : *ces acides racémisent moins facilement*.

3.2.2 Semantic results (using distributional clue)

Three kinds of dimensions have been identified and studied. The two first one concerns semantic phenomena strictly speaking:

- Synonymy : synonymy concerns cases in which a concept may correspond to two or more terms.

- Polysemy : polysemy corresponds to cases in which a term is associated to two or more concepts linkeded by a part of their meaning.

These two phenomena, synonymy and polysemy, may be observed either in the same domain or in two different domains. What is important to note is that, if they are often considered as critical problems when language is used only in its ability to be a vehicle for the provision of information, these phenomena are very interesting to observe when a field is under construction. They are sign of the creative power of the language and their fine-grained analysis is one of the best ways to spot creativity within the disciplines.

- Borrowing of a term ([18]) : concerns generally cases in which terms originate from another language. But in our study, the borrowing concerns rather terms which originate from another

discipline implicated in exobiology. In such cases, it may be difficult to know if the term retains its meaning or if it adopts a meaning related to the new discipline. If we consider that a discipline is based on a system, that is to say a structured representation, it is likely that the use of such terms will be adapted to the new discipline, that is to say, to the new point of view. Then, the meaning will be modified and will become more general or more precise.

All these phenomena show that meaning may change, specifically when several communities (specifically, scientific communities) are in contact and decide to collaborate. As we have said, such phenomena are highlighted by analysing contexts in which a form (term) appear, that is to say, by categorising the distribution of a term. But, in some cases, the writers are aware of such difficulties and express them. This point leads to the definition of the second dimension.

The second dimension concerns the fact that speakers are or are not aware of the semantic phenomena.

If experts are aware of certain semantic characteristics, it is possible to identify it because there are linguistic patterns (so-called metalinguistic patterns) that may be used to mean this awareness (for example, “as said in...”, “it is not the same sense in “, ...). In all cases, linguists may collect and estimate these phenomena and propose them to experts of all the disciplines concerned in order to initiate the basis of the discussion about the concepts in question.

The third dimension concerns the fact that semantic phenomena may or not lead to conflicts between experts. Language may be a “place of power” so the control of uses and definitions of terms may constitute an important issue for scientists. Some contexts show that the definition of certain concepts are fiercely defended.

So taking into account these three dimensions, we reach to twelve possibilities concerning the way terms work in a multidisciplinary domain. Here is a table summarizing these possibilities.

Table 2. The twelve categories of lexical phenomena in the corpus

	Aware	Not-aware	Controversial	Non controversial
Polysemy	x		x	
	x			x
		x	x	
		x		x
Synonymy	x		x	
	x			x
		x	x	
		x		x
Borrowing	x		x	
	x			x
		x	x	
		x		x

3.3 Examples

Here are some examples and the way they can be interpreted regarding the categorization presented previously.

Polysemy probably without awareness of the speakers.
The three extracts below contains the term (adjective in all the cases) *prebiotic*.

1- *Une atmosphère constituée d'azote moléculaire, de méthane et de vapeur d'eau constitue « la meilleure atmosphère prébiotique » i.e. le mélange gazeux le plus favorable à la synthèse des briques du vivant. (astronomie)*[b]

2- *Une des difficultés de la reconstitution de l'atmosphère prébiotique (avant l'apparition de la vie) réside dans l'impossibilité actuelle de dater les débuts de la vie sur Terre. (astronomie)*[c]

3- *La chimie prébiotique est une chimie organique en solution aqueuse, dans des conditions plausible de l'environnement primitif terrestre, conduisant à des composés d'intérêt biologique. (chimie)*[d]

A fine-grained analysis shows that *prebiotic* is not used exactly with the same meaning within the three examples. Note that the two first ones come from astronomy and the third, from chemistry.
The three examples contain what is considered as definitional patterns, more or less complete. In 3-, we encounter what is considered as an Aristotelian definition (In The Metaphysics).
[Definiendum = definiens + specificities] (with the definiens which is generally a generic term for the definiendum). In discourse, the copula = may expressed by *is a* [19]. So, from 3-, we can presume that *organical chemistry* is a generic term for *prebiotic chemistry*.
In 2-, there is also a definitional pattern (*before the beginning of life*) but less explicit because the copula is not as clearly expressed as in 3-. What can help to understand that the parenthesis concerns *prebiotic* is etymological elements: *pre-* and *biotic* from which we can infer the link with *before* and *life*.
In 1-, two kinds of patterns draw the linguists' attention. First, of course, *i.e* which announces a kind of paraphrase, but also the quotation marks around *la meilleure atmosphere prébiotique*. These quotation marks may be understood in two ways. They probably mean that this term is borrowed from another speaker (within or outside the discipline) but that it is not completely integrated by the writer of the extract. But they also enable to understand that the noun phrase following *i.e* concerns directly the term between quotation marks and that they are equivalent.

Not aware syonymy
The case of exoplanète (exoplanète) vs planète extrasolaire (extrasolar planet).
First of all, as we can see in table 3., Astronomy is the only discipline using the two terms *exoplanète* (exoplanet) and *planète extrasolaire* (extrasolar planet).

Table 3. Repartition of exoplanète and planète extra-solaire

	Astronomy	Geology	Chemistry	Biology
Exoplanète	19	0	3	0
Planète extrasolaire	12	1	0	3

Even before the study of contexts, we can suppose that these two terms are synonyms, using etymological clue: *exo-* and *extra-*, in greek for the first one and in latin for the second, mean "outside of". The term *planète extrasolaire* is more precise because it says that the exteriority

[b] traduction
[c] traduction
[d] traduction

concerns the sun. Its equivalence is confirmed by the two examples below in which these terms appear in very similar contexts.

1- *Signatures spectroscopiques de vie sur les exoplanètes.*
2- *Chercher la vie sur les planètes extrasolaires par la détection de raies d'oxygène dans leur spectre.*

Borrowing with awarness of speakers

The three examples below contain the adjective *inert.*

3- Les gaz rares des planètes (Ne, Ar, Kr, et Xe) sont chimiquement inertes. » (astronomie)
4- « Cependant, si cet appauvrissement est particulièrement marqué dans le cas des gaz rares, qui sont chimiquement inertes et donc peu retenus dans les silicates et le métal (…) » (géologie)
5- Une fois récupérés sous atmosphère inerte, ces produits sont soumis à différentes analyses afin de déterminer leurs propriétés optiques, leur solubilité dans différents solvants, leur structure moléculaire… » (chimie)

In the two first ones, *inert* is preceded by *chemically.* We can make the hypothesis that *chemically* has to be interpreted as: from a chemical point of view. So, writers (both in astronomy and geology) are aware that when they speak about the "inertia" of noble gas, they adopt a chemical point of view and they do not call into question this point (see the use of *then* in 8-). In 9, written by a chemistry expert, the same adjective is not preceded by *chemically*.

These examples show how the analysis of contexts may be partly systematized. Such studies may be long and they have to be confirmed or debated by domains experts. But they may be very useful in order to draw a picture of the situation at a moment in a multidisciplinary field.

4 Conclusion

In this chapter, the focus is put on the construction of the exobiology. Linguistics may play a role in such a construction by analyzing textual productions of the four disciplines implied in exobiology. By interpreting clues provided by tools, it is possible to show how the meaning is emerging, what constitutes the sign that the discipline itself is emerging. Twelve categories of linguistic phenomena have been identified. They take into account semantic phenomena such as polysemy, synonymy and borrowing but also the awareness (or not) of the writers and their possible consequences in relationship terms between disciplines (controversial or not). Real examples of the realization of these phenomena in the corpus of texts have been detailed. Some contexts, named linguistic patterns, may be directly interpreted as marking the awareness or the potential controversy. But in most cases, it is necessary to analyse in details what is called the distribution of a word, that is to say, all the contexts in which it appears.
In any case, these results are just propositions and they have to be discussed by exobiologists in order that they may build stable definitions, with a clear perspective their on scoping and adequacy.

References

1. T. Kuhn, *The Structure of Scientific Revolutions.* **University of Chicago Press** (1962)
2. A. Rey,*Essays on Terminology*, Amsterdam / Philadelphia, **John Benjamins Publishing Company** (1995)

3. G. Fourez, *La construction des sciences,* Bruxelles, De Boeck Université, (2002)
4. S. Atkins, J. Clear, N. Ostler, Corpus Design Criteria, Literary and Linguistic Computing, **7 (1)**, 1-16, (1992)
5. T. Mc Enery, A. Wilson. *Corpus Linguistics*. Edinburgh: **Edinburgh University Press**, (2004), (1st edition: 1996)
6. E. Tognini-Bonelli, *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins Publishing Company (2001)
7. Z. Harris, *A Theory of Language and Information: A Mathematical Approach*, Oxford University Press, (1991)
8. J. R. Firth, *Papers in Linguistics (1934-1951),* Oxford University Press, (1957)
9. M. A. K. Halliday, J.R. Martin, *Writing Science: Literacy and Discursive Power*. London, The Falmer Press (1993)
10. C. Gledhill, *Collocations in Science Writing,* Tübingen, Gunter Narr, 7-20, (2000)
11. J. Pearson, *Terms in Context*, Amsterdam and Philadelphia, **John Benjamins** (1998)
12. I. Stengers, J. Schlanger, *Les concepts scientifiques*, Paris, Gallimard, (1991)
13. G. Myers, *Writing Biology, texts in the Social Construction of Scientific Knowledge*. Madison, Wisconsin: **The University of Wisconsin Press** (1990)
14. F. Rastier, M. Cavazza, A. Abeillé, *Semantics for Descriptions*, Chicago, **Chicago University Press** (2002)
15. I. Prigogine, I. Stengers, *Order out of Chaos: Man's New Dialogue with Nature*. Toronto, New York, London, Sydney: **Bantam Books** (1984)
16. R. Temmerman, Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology, *Hermes* **18**, 51-90 (1997)
17. A. Condamines, N. Dehaut, Mise en œuvre des méthodes de la linguistique de corpus pour étudier les termes en situation d'innovation disciplinaire : le cas de l'exobiologie. *META*, **56(2)**, 266-283. (2011)
18. M. T. Cabré, *Terminology : theory, methods and applications*, Amsterdam and Philadelphia, **John Benjamins,** (1999)
19. A. Condamines, Corpus Analysis and Conceptual Relation Patterns, *Terminology*, 8-1 141-162 (2002)

[i] I thank Nathalie Dehaut for the data which she supplied me
[ii] This study was led thanks to the financial support of the CNRS and the CNES